Choice Models in Operations

Lecture 9: Estimation of MNL model

Instructor: Srikanth Jagabathula *Scribe: Xiao Lei*

At the end of last lecture, we discuss how to use Newton's method to estimation MNL model with single parameter, i.e., the utility is $U_j = \beta_j + \epsilon_j$. In this lecture, we first continue the discussion of Newton's method, and then discuss the necessary and sufficient condition for the existence of unique and bounded optimal estimation under the more general setting.

1 Discussion of Newton's Method

Recall that when the utility is $U_j = \beta_j + \epsilon_j$, we could use Newton's method to update $\underline{\beta}$, i.e., the Newton's update step is

$$\underline{\beta}^{(t+1)} = \underline{\beta}^{(t)} - \left(\nabla^2 l(\underline{\beta}^{(t)})\right)^{-1} \nabla l(\underline{\beta}^{(t)}), \tag{1}$$

where

$$\frac{\partial l(\underline{\beta})}{\partial \beta_k} = c_{k\cdot} - \sum_{r=1}^{L} c_{\cdot r} q_{kr}(\underline{\beta})$$

and

$$\frac{\partial^2 l(\underline{\beta})}{\partial \beta_k \beta_{k'}} = \begin{cases} \sum_{r=1}^{L} c_{\cdot r} q_{kr}(\underline{\beta}) q_{k'r}(\underline{\beta}) & \text{if } k \neq k', k' \in S_r \\ -\sum_{r=1}^{L} c_{\cdot r} q_{kr}(\underline{\beta})(1 - q_{kr}(\underline{\beta})) & \text{if } k = k' \end{cases} \tag{2}$$

However, we note that all the row sums of the Hessian are zero, which means that the Hessian is not invertible. Because of this, Newton's method cannot be directly applied. This is happening because the likelihood function is invariant under a constant translation,i.e.,

$$l(\underline{\beta} + c) = \sum_{t=1}^{T} \frac{e^{\beta_{j_t} + c}}{\sum_{i \in S_t} e^{\beta_i + c}} = \sum_{t=1}^{T} \frac{e^{\beta_{j_t}}}{\sum_{i \in S_t} e^{\beta_i}} = l(\underline{\beta}), \ \forall c \in \mathbb{R}.$$

Note that this only happens when the mean utility is constant. If $\beta_j = \underline{\gamma}^T x_j$, there is no transformation of $\underline{\gamma}$ such that $l(\underline{\gamma}) = l(g(\underline{\gamma}))$.

To deal with the issue of multiplicity of optima, we normalize the coefficient of one of the items, say item 1, i.e., $\beta_1 = 0$. Now the optimization problem becomes

$$\max_{\underline{\beta}} \ l(\underline{\beta})$$
$$\text{s.t. } \beta_1 = 0.$$

With this constraint, we just optimize over $\beta_2, \ldots, \beta_n$. The truncated Hessian is an $(n-1) \times (n-1)$ matrix with entries described in 2.

We end this section with the discussion of convergence rate. Newton's method is a second-order method with a quadratic rate of convergence, which is faster than the first-order method, which has a linear rate of convergence. However, the inverse of large matrices is computationally expensive.

2 MNL with Feature Vectors

We now switch to the more general setting where the products are described by features. In particular, we have observation $(j_t, S_t, (\underline{z}_i)_{i\in S_t}),\ t = 1, \ldots, T$, where $\underline{z}_i$ is the feature vector of item i. The MLE problem now becomes:

$$\begin{aligned}\max_{\underline{\beta}}\ l(\underline{\beta}) &= \max_{\underline{\beta}} \sum_{t=1}^{T} \log \frac{e^{\underline{\beta}^T \underline{z}_{j_t}}}{\sum_{i\in S_t} e^{\underline{\beta}^T \underline{z}_i}} \\ &= \max_{\underline{\beta}} \sum_{t=1}^{T} \left[\underline{\beta}^T \underline{z}_{j_t} - \log(\sum_{i\in S_t} e^{\underline{\beta}^T \underline{z}_i}) \right].\end{aligned}$$

We first show that $l(\underline{\beta})$ is globally concave in $\underline{\beta} \in \mathbb{R}^k$. To prove this, it is enough to show that the function $f(\underline{\beta}) = \log(\sum_{i\in S} e^{\underline{\beta}^T \underline{z}_i})$ is globally convex for $\forall \underline{z}_i \in \mathbb{R}^k$, which is equivalent to show that

$$f(\lambda \underline{\alpha} + (1-\lambda)\underline{\beta}) \le \lambda f(\underline{\alpha}) + (1-\lambda) f(\underline{\beta}),\ \lambda \in [0,1],\ \underline{\alpha}, \underline{\beta} \in \mathbb{R}^k.$$

The left hand side is

$$\begin{aligned}LHS &= \log\left(\sum_{i\in S} e^{(\lambda\underline{\alpha} + (1-\lambda)\underline{\beta})^T \underline{z}_i} \right) \\ &= \log\left(\sum_{i\in S} (e^{\underline{\alpha}^T \underline{z}_i})^{\lambda} (e^{\underline{\beta}^T \underline{z}_i})^{1-\lambda} \right).\end{aligned}$$

We now use Holder's inequality, which states that

$$\sum_{i=1}^{n} x_i y_i \le (\sum x_i^p)^{\frac{1}{p}} (\sum y_i^q)^{\frac{1}{q}},\ \text{s.t. } \frac{1}{p} + \frac{1}{q} = 1,\ p, q > 1,$$

and the equality occurs if and only if $x_i = c y_i, \forall i$ for some constant c. We now set $\lambda = 1/p,\ 1-\lambda = 1/q,\ x_i = (e^{\underline{\alpha}^T \underline{z}_i})^{\lambda},\ y_i = (e^{\underline{\beta}^T \underline{z}_i})^{1-\lambda}$, and applies Holder's inequality to the left hand side:

$$\begin{aligned}LHS &\le \log\left[\left(\sum \left((e^{\underline{\alpha}^T \underline{z}_i})^{\lambda} \right)^{\frac{1}{\lambda}} \right)^{\lambda} \left(\sum \left((e^{\underline{\beta}^T \underline{z}_i})^{1-\lambda} \right)^{\frac{1}{1-\lambda}} \right)^{1-\lambda} \right] \\ &= \log\left(\lambda \log\left(\sum e^{\underline{\alpha}^T \underline{z}_i} \right) + (1-\lambda) \log\left(\sum e^{\underline{\beta}^T \underline{z}_i} \right) \right) \\ &= \lambda f(\underline{\alpha}) + (1-\lambda) f(\underline{\beta}) = RHS,\end{aligned}$$

which proves the desired result.

Now suppose $\lambda \in (0,1)$, then it follows from Holder's inequality that $f(\lambda \underline{\alpha}+(1-\lambda)\underline{\beta}) \leq \lambda f(\underline{\alpha}) + (1-\lambda) f(\underline{\beta})$ if and only if

$$\begin{aligned} & \frac{(e^{\underline{\alpha}^T \underline{z}_i})^{\lambda}}{(e^{\underline{\beta}^T \underline{z}_i})^{(1-\lambda)}} = c, \ \forall i \in S \\ \longleftrightarrow & \ e^{(\lambda \underline{\alpha}+(1-\lambda)\underline{\beta})^T \underline{z}_i} = c, \ \forall i \in S \\ \longleftrightarrow & \ e^{\underline{\gamma}^T \underline{z}_i} = c, \ \forall i \in S, \text{where } \underline{\gamma} = \lambda \underline{\alpha} + (1-\lambda)\underline{\beta}. \end{aligned}$$

So $f(\underline{\beta})$ is strictly convex if and only if there doesn't exists $\underline{\gamma} \neq 0$ such that $\underline{\gamma}^T \underline{z}_i = c, \ \forall i \in S$.

3 Conditions for the Existence of Unique and Bounded Optimal Estimation

We firstly define a graph G with the n products as nodes, and there exists an edge from i to j if and only if i was chosen at least onece when j was also offered. In other words, we put edges from j_t to all $i \in S/\{j_t\}$.

Theorem 1 *The log-likelihood function $l(\underline{\beta})$ has a unique and bounded optimal solution if and only if*

(a) The graph G is strongly connected, i.e., there is a directed path between every pairs of distinct nodes.

(b) if $\underline{\gamma}(\underline{z}_i - \underline{z}_{j_t}) = 0, \ \forall i \in S_t/\{j_t\}, \ t = 1, \ldots, T$, then $\underline{\gamma} = 0$.

Proof of Sufficiency

We first argue that the log-likelihood function is strictly concave. Recall that

$$l(\underline{\beta}) = \sum_{t=1}^{T} \log \frac{e^{\underline{\beta}^T \underline{z}_{j_t}}}{\sum_{i \in S_t} e^{\underline{\beta}^T \underline{z}_i}},$$

where the functions in the summation is strictly concave because condition (b) ensures that only $\underline{\gamma} = 0$ makes $\underline{\gamma}^T \underline{z}_i = \underline{\gamma}^T \underline{z}_{j_t}, \ \forall t = 1, \ldots, T$, so by the proof in section 2 we know that $l(\underline{\beta})$ must be strictly concave.

Now we want to show that the optimal solution is bounded. For that, we argue that the optimal solution belongs to bounded ball.

We proceed as follows: define the unit sphere $s_k = \{\underline{\gamma} \in \mathbb{R}^k : ||r||_2 = 1\}$. Let $b(\underline{\gamma}) = \max_{i \in S_t/\{j_t\}} \underline{\gamma}^T (\underline{z}_i - \underline{z}_{j_t})$ for $t = 1, \ldots, T$. We claim that for each $\underline{\gamma} \in \mathbb{R}^k$, there exists at least one pair $i \in S_t$ and j_t such that $\underline{\gamma}^T (\underline{z}_i - \underline{z}_{j_t}) > 0$. (We prove this claim later)

Consider the following:

$$\begin{aligned} l(\underline{\beta}) &= \sum_{t=1}^{T} \left[-\log \left(\sum_{i \in S_t} e^{\underline{\beta}^T (\underline{z}_i - \underline{z}_{j_t})} \right) \right] \\ &= -\sum_{t=1}^{T} \left[\log \left(\sum_{i \in S_t} e^{\underline{\gamma}^T (\underline{z}_i - \underline{z}_{j_t}) \cdot ||\underline{\beta}||} \right) \right], \end{aligned}$$

where $\underline{\gamma} = \underline{\beta}/||\underline{\beta}|| \in s_k$.

It follows from our claim that there exist i^*, j_t^* such that $\underline{\gamma}^T(\underline{z}_{i^*} - \underline{z}_{j_t}) > 0$ for some t. We can now write

$$\log \left(\sum_{i \in S_t} e^{\underline{\gamma}^T (\underline{z}_i - \underline{z}_{j_t}) \cdot ||\beta||} \right) \geq \log e^{\underline{\gamma}^T (\underline{z}_{i^*} - \underline{z}_{j_t^*}) \cdot ||\underline{\beta}||} = \underline{\gamma}^T (\underline{z}_{i^*} - \underline{z}_{j_t^*}) \cdot ||\underline{\beta}||.$$

It thus follows that

$$\begin{aligned} l(\underline{\beta}) &\leq -\underline{\gamma}^T (\underline{z}_{i^*} - \underline{z}_{j_t^*}) ||\underline{\beta}|| \\ &\leq -b^* ||\underline{\beta}||, \end{aligned}$$

where $b^* = \min_{\underline{\gamma} \in s_k} b(\underline{\gamma})$. Note that b^* is well defined since s_k is compact. We will argue that $b^* > 0$ later.

Now choose $D = \{\underline{\beta} \in \mathbb{R}^k : ||\underline{\beta}|| \leq -l(\underline{0})/b^*\}$. Then for any $\underline{\beta} \notin D$,

$$\begin{aligned} l(\underline{\beta}) &\leq -b^* ||\underline{\beta}|| \\ &< -b^* (-\frac{0}{\underline{\beta}}) \\ &= l(0). \end{aligned}$$

Because $\underline{0} \in D$, the optimal solution must belong to D. We can now taking the maximum over a compact set D, the maximum is always achieved.

We now need to show that $b^* > 0$. It is sufficient to show that $b(\underline{\gamma}) > 0$ for all $\underline{\gamma}$. (The proof is left.)

We now argue our claim that given any $\underline{\gamma} \neq 0$, there exists a pair of i, j_t, such that $\underline{\gamma}^T(\underline{z}_i - \underline{z}_{j_t}) > 0$. Suppose this is not true, i.e., there exists a $\underline{\gamma} \neq 0$ such that $\underline{\gamma}^T(\underline{z}_i - \underline{z}_{j_t}) \leq 0$, $\forall i \in S_t/\{j_t\}$, $t = 1, \ldots, T$.

Now consider any two items $k \neq k'$. Because G is strongly connected, there exist directed paths $k \to l_1 \to l_2 \to \ldots \to l_m \to k'$ and $k' \to l_1' \to l_2' \to \ldots \to l_m' \to k$. This implies that

$$\underline{\gamma}^T(\underline{z}_{l_1} - \underline{z}_k) \leq 0,\ \underline{\gamma}^T(\underline{z}_{l_2} - \underline{z}_{l_1}) \leq 0, \ldots,\ \underline{\gamma}^T(\underline{z}_{k'} - \underline{z}_{l_m}) \leq 0$$

and

$$\underline{\gamma}^T(\underline{z}_{l_1'} - \underline{z}_{k'}) \leq 0,\ \underline{\gamma}^T(\underline{z}_{l_2'} - \underline{z}_{l_1'}) \leq 0, \ldots,\ \underline{\gamma}^T(\underline{z}_k - \underline{z}_{l_m'}) \leq 0.$$

Summing along each path, we have

$$\underline{\gamma}^T(\underline{z}_k - \underline{z}_{k'}) \leq 0, \underline{\gamma}^T(\underline{z}_{k'} - \underline{z}_k) \leq 0,$$

which implies that $\underline{\gamma}^T(\underline{z}_k - \underline{z}_{k'}) = 0$. So either $\gamma = 0$ or $\underline{z}_{k'} - \underline{z}_k = 0$, $\forall k \neq k'$. The latter one does not hold for general case, so $\underline{\gamma}$ has to be zero. But we assume $\underline{\gamma} \neq 0$, this contradicts to condition (b) that $\underline{\gamma}^T(\underline{z}_i - \underline{z}_{j_t}) = 0$ only when $\underline{\gamma} = 0$.

Proof of Necessity

Suppose condition (a) is violated, i.e., there exists $i.j$ such that there is no directed path from i to j. Let A denote the set of nodes that can be reached from i. Clearly $j \notin A$. Let's also include i in A. Let A^c denote the nodes that are not in A.

We now consider 2 cases.

1. Suppose there is no path from A^c to A. For all t such that $i \in S_t$, we must have that $S_t \subset A$. Similarly, if $i \notin S_t$, $S_t \subset A^c$. Let's assume that we have specific coefficients present for each product. In particular, let $\underline{z}_k = [\underline{e}_k | \underline{y}_k]$, where $\underline{e}_k$ is the unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$. Let α_k denote the dummy corresponding to item k. We can set $\alpha'_k = \alpha_k$ for all $k \in A$ and $\alpha'_k = \alpha_k + c$ for all $k \in A^c$. This would create a new coefficient vector with the same log-likelihood value. Hence the optimal solution will not be unique.

2. Suppose we have a path from A^c to A. Then set $\alpha'_k = \alpha_k$ for all $k \in A$ and $\alpha'_k = \alpha_k + c$ for all $k \in A^c$. As above, if $S \subset A$ or $S \subset A^c$ the log-likelihood value does not change. Now consider the case the S intersects both A and A^c, then it must be that $j_t \in A^c$, therefore the likelihood of the observation (j_t, S_t) becomes

$$\frac{e^{\alpha_{j_t} + c + \underline{\gamma}^T \underline{y}_{j_t}}}{\sum_{k \in S_t \cap A^c} e^{\alpha_k + c + \underline{\gamma}^T \underline{y}_k} + \sum_{k \in S_t \cap A} e^{\alpha_k + \underline{\gamma}^T y_{j_t}}},$$

because $x \to \frac{x}{1+x}$ is increasing in x, the above is strictly increasing in c. As $c \to \infty$, the log-likelihood will keep increasing. As a result, the optimal solution will be unbounded.

The necessity of condition (b) is left.